Quia Survey: Variables

Age	# of colleges applying to
Gender	years w/ cell phone
Hair Color	siblings
Eye Color	death penalty
Grade (10,11,12)	last math class
Height	middle school
zip code	Age (months)
first 3 Cell #'s	AP course before?
sum of cell digits	pizza chain
SATM & SATV	AP Exam in may?

AP STAT: CHAPTER 3 **CATEGORICAL DATA**

MAKE A PICTURE!

First, create a frequency table

Example: number of students at CB South in each grade: (total 1601)

Grade	TOTAL
10	534
11	552
12	515

rounding 3-4 dec. places

Proportion = decimal Ex: 534/1601 = 0.3335

Percent = % Ex: 534/1601 = 33.35%

Frequency = # of things Ex: 534 sophomores

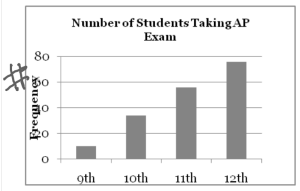
Relative frequency = % of things Ex: 33.35% sophomores

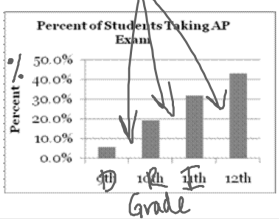
Distribution (of a variable)- shows values of the variable & how often the sample takes each value

Examples: Bar chart, pie chart, histogram, stemplot, etc.

Categorical Distributions:

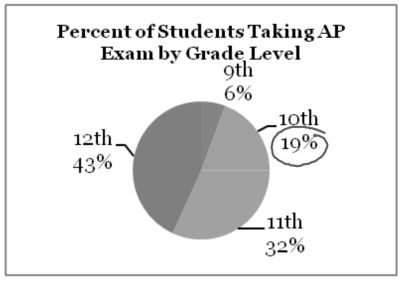
1. Bar Chart





spaces

Pie charts



D =
R =
I =

3. Contingency tables (aka Two-Way tables)

	Frosh	Soph	Junior	Senior	Total
Male					
Female					
Total					n

Identify:

- Row variable = GENDER (2 values: Male/Female)
- Column variable = GRADE (4 values: Fr, So, Jr, Sen)
- Values of the variable = the different rows/columns
- Total (n) = bottom right of chart n = sample size
- # of Cells (doesn't count totals) 8
- Totals (margins)

Example: Hospitals

Survived \ Hospital	Hospital A	Hospital B	total
Died	63	16	79
Survived	2037	784	2821
total	2100	800	2900

* What percent of people died?

79/2900 = 2.72% = 0.0272 = P(D) = P(Died)

Notation:

Probability: P(____) =

Given/Of:

And: P(S ∩ A)

Or: P(S ∪ A)

* **Of** those people that went to Hospital A, what percent died?

63/2100 = P(D | A)

	Hospital A	Hospital B	
Died	63	16	79
Survived	2037	784	2821
	2100	800	2900

* **Given that** someone went to Hospital B, what's the chance that they died? $P(D \mid B) = 16/800 = 2\%$

* **Of** those people who died, what percent went to Hospital A?

$P(A \mid D) = 63/79 = 0.7975$

* What percent of people died **and** went to Hospital B?

$P(D \cap B) = 16/2900 = 0.0055$

* What percent of people survived **or** went to Hospital A?

$P(S \cup A) = 2884/2900 = 99.45\%$

2 types of Distributions for Categorical Variables

1) MARGINAL DISTRIBUTIONS

- How to make: convert totals into %'s
- *Example: Hair color vs. Gender*

	Brown	Blonde	Black	Red	Total
MALE	26	24	10	3	63
FEMALE	20	35	12	6	73
TOTALs	46	59	22	9	136

* Marginal distrib. of HAIR COLOR:

Brown = 46/136 = 33.82%
Blonde = 59/136 = 43.38%
Black = 22/136 = 16.18%
Red = 9/136 = 6.62%

* Marginal distribution of GENDER:

Male = 46.32%
Female = 53.68%

* Visually: BAR GRAPH

2) CONDITIONAL DISTRIBUTIONS

- Look at … one variable
- Then look at … each value of the variable
- Break down … each value into its pieces
- ALWAYS … in %
- Example: Hair Color vs. Gender

$P(\ |\ )$

Find the conditional distrib. of HAIR COLOR:

	Brown	Blonde	Black	Red	Total
MALE	26	24	10	3	63
FEMALE	20	35	12	6	73
TOTALs	46	59	22	9	136

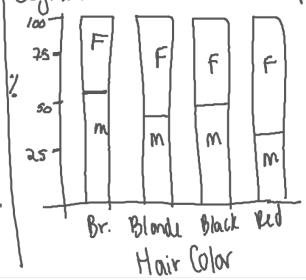
Brown
M = 26/46 = 56.52%
F = 20/46 = 43.48%

Blonde
M = 40.68%
F = 59.32%

Black
M = 45.45%
F = 54.55%

Red
M = 33.33%
F = 66.67%

Segmented or Stacked Bar Graph



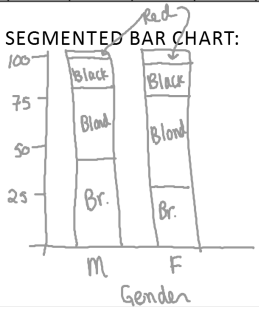
SEGMENTED BAR CHART:

Find the conditional distrib. of GENDER:

	Brown	Blonde	Black	Red	Total
MALE	26	24	10	3	63
FEMALE	20	35	12	6	73
TOTALs	46	59	22	9	136

	M	F
Br.	41.27%	27.4%
Blon	38.10%	47.95%
Bla	15.87%	16.44%
R	4.76%	8.22%

SEGMENTED BAR CHART:



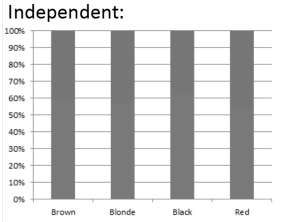
Independence:

- When one variable has no effect on the another variable

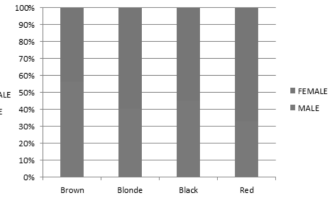
relationship, association

How do we tell independence? When the conditional distribution looks the same throughout all values of the variable (when all bars of the conditional look approx. the same). The same = each category within 5%

Independent:



Dependent:



Try worksheet 3A on your own!
Feel free to work with someone else

AP Stat- worksheet 3A- Categorical Variables practice

In a survey of adult Americans, people were asked to indicate their **age** and to categorize their **political preference** (liberal, moderate, conservative). The results are as follows:

	Liberal	Moderate	Conservative	*Total*
under 30	83	140	73	*296*
30 - 50	119	280	161	*560*
over 50	88	284	214	*586*
total	*290*	*704*	*448*	*1442*

1. What are the row and column variables?
2. What percent of Liberals are under 30?
3. Of those over 50, what percent are Liberals?
4. Of those that are moderates, what percent are 30-50?
5. What percent of respondents are moderate and under 30?

P( )

6. Calculate the ***marginal distribution*** for the ***AGE variable***. Write these down. Then make a bar graph of the marginal distribution for age.
7. Calculate the ***marginal distribution*** for the ***PREFERENCE variable***. Write these down. Then make a bar graph of this marginal distribution.
8. Calculate the ***conditional distribution*** of the ***AGE variable***. Write these down. Then make a segmented bar graph of this marginal distribution.
9. Calculate the ***conditional distribution*** of the ***PREFERENCE variable***. Write these down. Then make a segmented bar graph of this marginal distribution.
10. Are the two variables independent? Justify.

1) Row var = AGE Column var = POLITICAL PREFERENCE

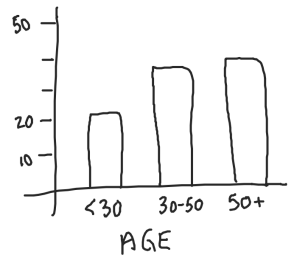
2) P(<30|L) = 83/290 = 28.6%

3) P(L|50+) = 88/586 = 15%

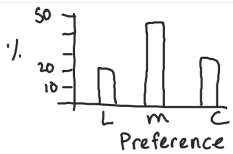
4) P(30-50|M) = 280/704 = 39.8%

5) P(M ∩ <30) = 140/1442 = 9.7%

6) AGE
<30 = 20.53%
30-50 = 38.83%
50+ = 40.64%

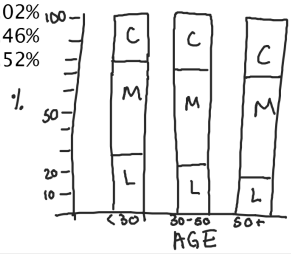


7) PREFERENCE
L = 20.11%
M = 48.82%
C = 31.07%



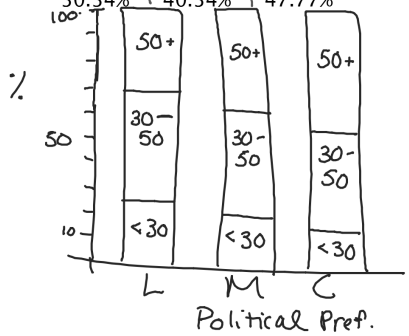
8)

	<30	30-50	50+
L	28.04%	21.25%	15.02%
M	47.3%	50%	48.46%
C	24.66%	28.75%	36.52%



9)

	L	M	C
<30	28.62%	19.89%	16.29%
30-50	41.03%	39.77%	35.94%
50+	30.34%	40.34%	47.77%



10. Independent:

No, they are not independent of each other. They are dependent. Age has an affect on political affilitation.

Looking at the calculated percentages and the segmented bar graph, the percentages of under 30 year olds and over 50 year olds in each of the 3 political classes are more than 5% different.

Ex: Under 30 year olds are 28.62% liberal, 19.89% moderate, and 16.29% conservative.

AP Stat- worksheet 3B- Categorical Variable practice

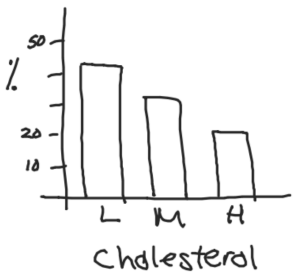
A 4-year study reported in *The New York Times*, on men more than 70 years old analyzed blood cholesterol and noted how many men with different cholesterol levels suffered nonfatal or fatal heart attacks.

	Low cholesterol	Medium cholesterol	High cholesterol	
Nonfatal heart attacks	29	17	18	64
Fatal heart attacks	19	20	9	48
	48	37	27	112

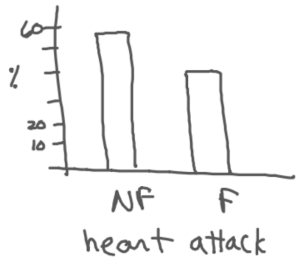
a. Calculate the marginal distribution for cholesterol level and make a bar graph.
b. Calculate the marginal distribution for severity of heart attack and make a bar graph.
c. Calculate three conditional distributions for the three levels of cholesterol and make a stacked bar graph.
d. Calculate the conditional distributions for the type of heart attack and make a stacked bar graph.

e. Independent?

(a) low: 42.9%
med: 33.0%
high: 24.1%

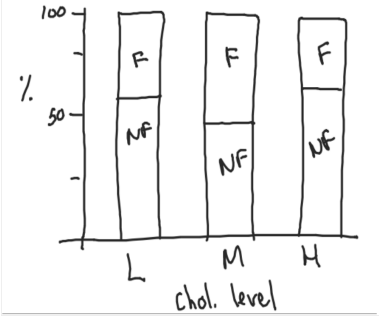


(b) NF: 57.1%
F: 42.9%



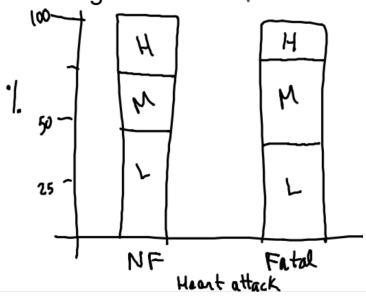
(c)

	low	med	high
NF	60.4%	45.9%	66.7%
F	39.6%	54.1%	33.3%



(d)

	NF	Fatal
Low:	45.3%	39.6%
Med:	26.6%	41.7%
High:	28.1%	18.8%



(e) Independence?

No, they are not independent, they are dependent. There is a relationship between type of heart attack and cholesterol levels

There is more than a 5% difference in each level of cholesterol for the non fatal heart attacks.

Try page 42 #33

Are the two variables dependent or independent? Justify.